# REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188



| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| 26-03-2013 | Final Report | 15-Aug-2010 - 14-May-2011 |

**4. TITLE AND SUBTITLE**
Final Report for ARMY STIR Grant W911NF-10-1-0360: Inferring Implicit Human Social Network Structure from Multi-modal Data

**5a. CONTRACT NUMBER**
W911NF-10-1-0360

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
611102

**6. AUTHORS**
Sanjay Shakkottai, Sujay Sanghavi

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAMES AND ADDRESSES**
University of Texas at Austin
The University of Texas at Austin
101 East 27th Street
Austin, TX 78712 -1539

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

**10. SPONSOR/MONITOR'S ACRONYM(S)**
ARO

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
58358-NS-II.2

**12. DISTRIBUTION AVAILIBILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**13. SUPPLEMENTARY NOTES**
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

**14. ABSTRACT**
Markov Random Fields (MRFs), a.k.a. Graphical Models, serve as popular models for networks in the social and biological sciences, as well as communications and signal processing. A central problem is one of structure learning or model selection: given samples from the MRF, determine the graph structure of the underlying distribution. When the MRF is not Gaussian (e.g. the Ising model) and contains cycles, structure learning is known to be NP hard even with infinite samples. Existing approaches typically focus either on specific parametric classes

**15. SUBJECT TERMS**
graphical models, greedy learning

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 15. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Sanjay Shakkottai |
| UU | UU | UU | UU | | 19b. TELEPHONE NUMBER 512-471-5376 |

**Report Title**

Final Report for ARMY STIR Grant W911NF-10-1-0360: Inferring Implicit Human Social Network Structure from Multi-modal Data


**ABSTRACT**

Markov Random Fields (MRFs), a.k.a. Graphical Models, serve as popular models for networks in the social and biological sciences, as well as communications and signal processing. A central problem is one of structure learning or model selection: given samples from the MRF, determine the graph structure of the underlying distribution. When the MRF is not Gaussian (e.g. the Ising model) and contains cycles, structure learning is known to be NP hard even with infinite samples. Existing approaches typically focus either on specific parametric classes of models, or on the sub-class of graphs with bounded degree; the complexity of many of these methods grows quickly in the degree bound. We develop a simple new 'greedy' algorithm for learning the structure of graphical models of discrete random variables. It learns the Markov neighborhood of a node by sequentially adding to it the node that produces the highest reduction in conditional entropy. In our work, we provide a general sufficient condition for exact structure recovery (under conditions on the degree/girth/correlation decay), and study its sample and computational complexity. We then consider its implications for the Ising model, for which we establish a self-contained condition for exact structure recovery.


---

## Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

**(a) Papers published in peer-reviewed journals (N/A for none)**

| Received | Paper |
|---|---|

**TOTAL:**

**Number of Papers published in peer-reviewed journals:**

---

**(b) Papers published in non-peer-reviewed journals (N/A for none)**

| Received | Paper |
|---|---|

**TOTAL:**

**Number of Papers published in non peer-reviewed journals:**

---

**(c) Presentations**

**Number of Presentations:** 0.00

## Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

| Received | Paper |
|---|---|

**TOTAL:**

**Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):**

## Peer-Reviewed Conference Proceeding publications (other than abstracts):

| Received | Paper | |
|---|---|---|
| 03/14/2013 | 1.00 | Praneeth Netrapalli, Siddhartha Banerjee, Sujay Sanghavi, Sanjay Shakkottai. Greedy learning of Markov network structure,<br>2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton). 2010/09/28 01:00:00, Monticello, IL, USA. : , |

**TOTAL:** **1**

**Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):**

## (d) Manuscripts

| Received | Paper |
|---|---|

**TOTAL:**

**Number of Manuscripts:**

## Books

| Received | Paper |
|---|---|

**TOTAL:**

## Patents Submitted

## Patents Awarded

## Awards

## Graduate Students

| NAME | PERCENT_SUPPORTED | Discipline |
|---|---|---|
| Siddhartha Banerjee | 0.75 | |
| Praneeth Netrapalli | 0.50 | |
| **FTE Equivalent:** | **1.25** | |
| **Total Number:** | **2** | |

## Names of Post Doctorates

| NAME | PERCENT_SUPPORTED |
|---|---|
| **FTE Equivalent:** | |
| **Total Number:** | |

## Names of Faculty Supported

| NAME | PERCENT_SUPPORTED |
|---|---|
| **FTE Equivalent:** | |
| **Total Number:** | |

## Names of Under Graduate students supported

| NAME | PERCENT_SUPPORTED |
|---|---|
| **FTE Equivalent:** | |
| **Total Number:** | |

## Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

| | |
|---|---|
| The number of undergraduates funded by this agreement who graduated during this period: ..... | 0.00 |
| The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... | 0.00 |
| The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... | 0.00 |
| Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale): ..... | 0.00 |
| Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering: ..... | 0.00 |
| The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense ..... | 0.00 |
| The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: ..... | 0.00 |

## Names of Personnel receiving masters degrees

NAME

**Total Number:**

## Names of personnel receiving PHDs

NAME

**Total Number:**

## Names of other research staff

NAME PERCENT_SUPPORTED

**FTE Equivalent:**

**Total Number:**

## Sub Contractors (DD882)

## Inventions (DD882)

## Scientific Progress

See Attachment

## Technology Transfer

**Final Report for ARMY STIR Grant W911NF-10-1-0360: Inferring Implicit Human Social Network Structure from Multi-modal Data**

**Summary:**

This proposal was a 9-month STIR that explored the development of algorithms with provable guarantees for Markov Random Fields (graphical models) structure learning, with applications to social networks.

Markov Random Fields (MRFs), a.k.a. Graphical Models, serve as popular models for networks in the social and biological sciences, as well as communications and signal processing. A central problem is one of structure learning or model selection: given samples from the MRF, determine the graph structure of the underlying distribution. When the MRF is not Gaussian (e.g. the Ising model) and contains cycles, structure learning is known to be NP hard even with infinite samples. Existing approaches typically focus either on specific parametric classes of models, or on the sub-class of graphs with bounded degree; the complexity of many of these methods grows quickly in the degree bound. We develop a simple new 'greedy' algorithm for learning the structure of graphical models of discrete random variables. It learns the Markov neighborhood of a node by sequentially adding to it the node that produces the highest reduction in conditional entropy.

In our work, we provide a general sufficient condition for exact structure recovery (under conditions on the degree/girth/correlation decay), and study its sample and computational complexity. We then consider its implications for the Ising model, for which we establish a self-contained condition for exact structure recovery.

Further, we present numerical results that highlight the applicability of this approach for social network relationship learning. The results summarized in this document are elaborated in much greater technical depth in the included technical report. An early version of some of the results that resulted from this STIR are presented in:

*P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai. Greedy learning of Markov network structure. In 48th Annual Allerton Conference on Communication, Control and Computing, pages 1295 –1302, Sept. 29 - Oct. 1 2010.*

**Outline of Results in the Technical Report:**

1. Algorithm: A greedy algorithm is proposed for learning (pp. 7) that takes as input, samples from the MRF and outputs the graph structure. This is done in a sequential and greedy manner, where a node at each time adds a single additional node as a neighbor that most decreases its conditional entropy conditioned its neighborhood.

2. Result: Under non-degeneracy, degree bounds, and correlation decay assumptions, we show that this algorithm recovers the correct graphical model structure. We further show that an Ising model (with some assumptions) satisfy these conditions, see Theorem 7, pp.

14 in the included technical report.

3. We study the applicability of the algorithm for the well-known senator voting records dataset (see pp. 16, and O. Banerjee et. al. pp. 18), and demonstrate that the algorithm recovers our intuition on voting patterns (e.g., same state senators tend to vote together, same part senators tend to vote together); however, the algorithm does so purely based on the data and with no "side information" on political knowledge. Further, the algorithm reveals the senators who tend to vote "across the aisle". See plot below, and also pp. 20 of the attached technical report.

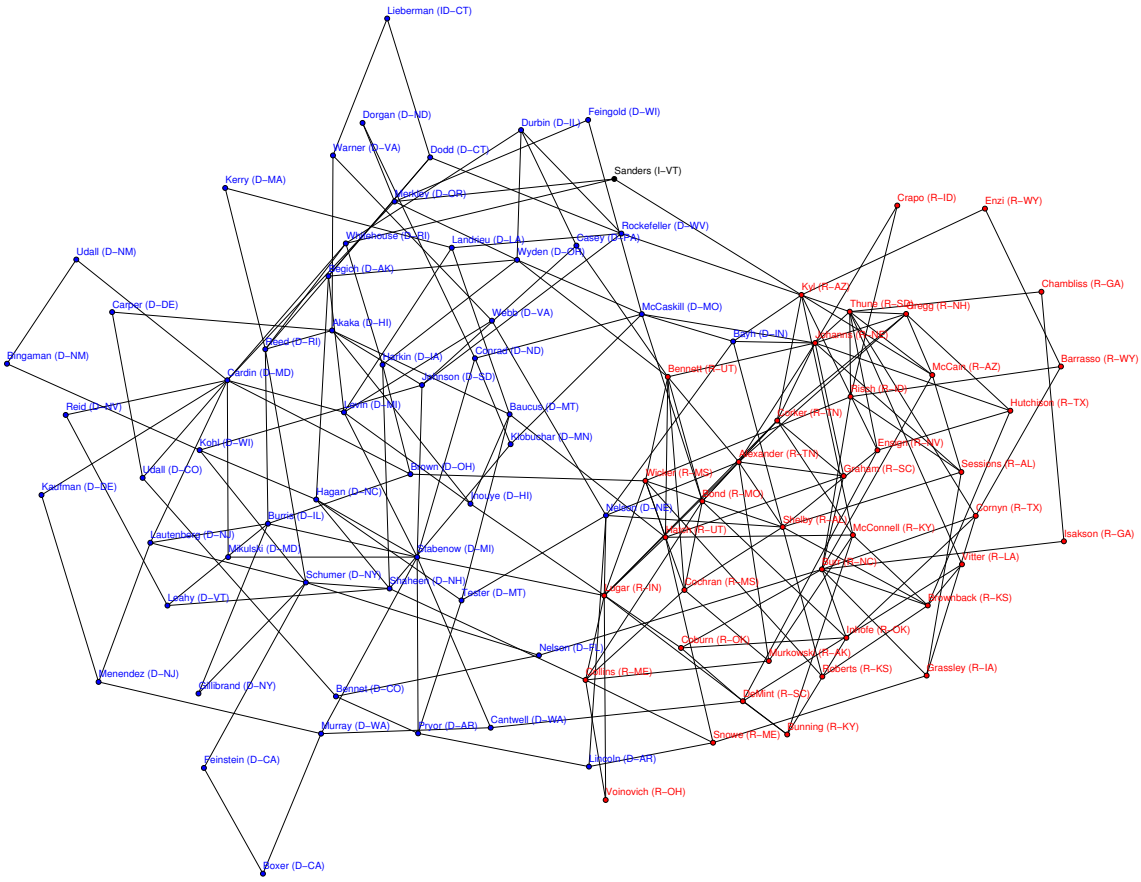



Figure 1: Following the approach in Banerjee et al., 2008, we present an application of our algorithm to model senator interaction graph using the senate voting records. Blue nodes represent democrats, red nodes represent republicans and black node represents an independent. We use a value of 0.05 for ε in the algorithm. We can make some preliminary observations from the graph. Most of the democrats are connected to other democrats and most of the republicans are connected to other republicans (in particular, the number of edges between democrats and republicans is approximately 0.1 fraction of the total number of edges). The senate minority leader, McConnell, is well connected to other republicans where as the senate majority leader, Reid, is not well connected to other democrats. Sanders and Lieberman, both of who caucus with democrats have more edges to democrats than to republicans. We use the graph drawing algorithm of Kamada and Kawai to render the graph (Kamada and Kawai, 1989).

# Greedy Learning of Markov Network Structure *

**Praneeth Netrapalli** PRANEETHN@UTEXAS.EDU
**Siddhartha Banerjee** SBANERJEE@MAIL.UTEXAS.EDU
**Sujay Sanghavi** SANGHAVI@MAIL.UTEXAS.EDU
**Sanjay Shakkottai** SHAKKOTT@MAIL.UTEXAS.EDU
*Department of Electrical and Computer Engineering*
*The University of Texas at Austin*
*Austin, TX 78712, USA*

**Editor:**

## Abstract

Markov Random Fields (MRFs), a.k.a. Graphical Models, serve as popular models for networks in the social and biological sciences, as well as communications and signal processing. A central problem is one of structure learning or model selection: given samples from the MRF, determine the graph structure of the underlying distribution. When the MRF is not Gaussian (e.g. the Ising model) and contains cycles, structure learning is known to be NP hard even with infinite samples. Existing approaches typically focus either on specific parametric classes of models, or on the sub-class of graphs with bounded degree; the complexity of many of these methods grows quickly in the degree bound. We develop a simple new 'greedy' algorithm for learning the structure of graphical models of discrete random variables. It learns the Markov neighborhood of a node by sequentially adding to it the node that produces the highest reduction in conditional entropy. We provide a general sufficient condition for exact structure recovery (under conditions on the degree/girth/correlation decay), and study its sample and computational complexity. We then consider its implications for the Ising model, for which we establish a self-contained condition for exact structure recovery.

## 1. Introduction

Markov Random Fields (MRF) are undirected graphical models which are used to encode conditional independence relations between random variables. At a more abstract level, a graphical model captures the dependencies between a collection of entities. Thus the nodes of a graphical model may represent people, genes, languages, processes, etc., while the graphical model illustrates certain conditional dependencies among them (for example, influence in a social network, physiological functionality in genetic networks, etc.). Often the knowledge of the underlying graph is not available beforehand, but must be inferred from certain observations of the system. In mathematical terms, these observations correspond to samples drawn from the underlying distribution. Thus, the core task of structure learning is that of inferring conditional dependencies between random variables from i.i.d samples drawn from their joint distribution. The importance of the MRF in understanding

∗. The results in this paper were presented in (Netrapalli et al., 2010) without proofs of the theorems. This paper includes all the proofs along with simulations.

the underlying system makes structure learning an important primitive for studying such systems.

More specifically, an MRF is an undirected graph $G(V, E)$, where the vertex set $V = \{v_1, v_2, \ldots, v_p\}$ corresponds to a $p$-dimensional random variable $X = \{X_1, X_2, \ldots, X_p\}$ (whereby each vertex $i$ is associated with variable $X_i$), and the edges encode the conditional dependencies between the random variables (this is explained in detail in Section 2). A structure learning algorithm takes as input, samples drawn from the distribution of $X$, and outputs an estimate $\hat{G}$ of the underlying MRF. There are three primary yardsticks for a structure learning algorithm:- correctness, sample complexity and computational complexity. The three are interdependent, and in a sense an ideal structure learning algorithm is one which can learn any underlying graph on the nodes with high probability (or with probability of error less than some given $\delta$, analogous to the PAC model of learning) with associated sample complexity and computational complexity polynomial in $p$ and $\frac{1}{\delta}$. However, it is known that the general structure learning problem is a difficult problem, both in terms of sample complexity (Santhanam and Wainwright, 2009; Bento and Montanari, 2009) and computational complexity (Srebro, 2003; Bogdanov et al., 2008). Inspite of this, the practical importance of the problem has motivated a lot of work in this topic, and there are several approaches in the literature that, although not optimal, perform well (both in practice, and also theoretically) under some stronger constraints on the problem.

There are two fundamental ways to perform structure learning, corresponding to two different interpretations of a graphical model. Under certain conditions (given by the Hammersley-Clifford theorem (Wainwright and Jordan, 2008)), the conditional independence view of a graphical model leads to a factorization of the joint probability mass function (or density) according to the cliques of the graph. *Parameter estimation techniques* (Ravikumar et al., 2010; Banerjee et al., 2008) utilize such a factorization of the distribution to learn the underlying graph. These techniques assume a certain form of the potential function, and thereby relate the structure learning problem to one of finding a sparse maximum likelihood estimator of a distribution from its samples. On the other hand, algorithms based on learning conditional independence relations between the variables, which we refer to as *comparison tests*, are potential agnostic, i.e., they do not need knowledge of the underlying parametrization to learn the graph. These methods are based on comparing all possible neighborhoods of a node to find one which has the 'maximum influence' on the node. In both cases, in order to learn the underlying graph accurately and efficiently, the algorithms need some assumptions on the underlying distribution and graph structure. There are several existing comparison test based methods (Chow and Liu, 1968; Abbeel et al., 2006; Bresler et al., 2008; Anandkumar and Tan, 2011a,b), each with associated conditions under which they can learn the graph correctly.

In addition to the difference in underlying assumptions, there is another fundamental difference in the philosophy of the two approaches. The parameter estimation techniques tend to be 'bottom-up' approaches, whereby the algorithm is proposed first, based on some intuition regarding the system, and then subsequently it is analyzed and conditions are found for correctness and efficiency. On the other hand, the comparison-test techniques in literature tend to be designed with the aim of achieving some correctness requirements. As a result, comparison-test algorithms usually involve a computationally expensive search over all potential neighborhoods of a node, and this increases their computational complexity.

In addition, although these algorithms make no assumptions on the parametrization of the distribution, they need to assume some properties of the graph in order to succeed (for example, the algorithm of Bresler et. al. (Bresler et al., 2008) needs to know the maximum degree of the graph in order to learn it). Our contribution in this work is to propose a simple 'greedy', comparison-test based algorithm for learning MRF structure. As in any sub-optimal greedy algorithm, we can not always guarantee correctness, but are guaranteed low computational complexity. However, we are able to provide general sufficient conditions for the success of the algorithm for any graphical model, and show that these conditions are in fact satisfied by one specific graphical model of significance in literature: the pairwise symmetric binary model, or the Ising model.

Greedy comparison-tests for exact structure learning are however not completely new, and in fact one of the early successes in the field was in the form of a greedy algorithm. In their seminal paper, Chow and Liu (Chow and Liu, 1968) showed that if the MRF was a tree, then it could be learnt by a simple maximum spanning tree algorithm. However their method is crucially dependent on the underlying graph being a spanning tree (although recent results (Tan et al., 2010) have shown how it can be modified to learn general acyclic graphs), and fails as soon as the graph has loops. Our algorithm, in some sense, generalizes the Chow and Liu algorithm to a richer class of graphs. This is in spirit similar to the manner in which loopy belief propagation extends the dynamic programming paradigm from trees to loopy graphs. One notes however that unlike the Chow and Liu algorithm which searches for a globally optimal graph, ours is a locally greedy algorithm, whereby we learn the neighborhood of each node separately in a greedy manner.

The remaining sections are organized as follows. In Section 2, we review graphical models and some results from information theory, and set up the structure learning problem. Our new structure learning algorithm, GreedyAlgorithm($\epsilon$), is given in Section 3. Next, in Section 4, we develop a sufficient condition for the correctness of the algorithm for general graphs. To demonstrate the applicability of this condition, we translate it into equivalent conditions for learning an Ising model in Section 5. We present simulation results evaluating our algorithm in Section 6. We discuss future work and conclude in Section 7. The proofs of theorems are in the Appendix.

## 2. Preliminaries

In this section, we formally define a graphical model and set up the structure learning algorithm. In addition, as a foreshadow to our structure learning algorithm, we define conditional entropy, and state some of its properties which we use later. We also define a notion of 'empirical' conditional entropy which we later use as our test function, and state an important lemma from information theory that helps relate empirical entropy and empirical measures. For more details regarding graphical models, refer to (Wainwright and Jordan, 2008), and for the information theoretic definitions, refer to (Cover and Thomas, 2006).

First we establish some notation that we use throughout. We assume in this paper that the random vector $X$ whose graph we are trying to learn is discrete valued. More specifically, we assume that $X$ is an $n$-dimensional random vector $\{X_1, X_2, \ldots, X_n\}$, where each component $X_i$ of $X$ takes values in a finite set $\mathcal{X}$. We use the shorthand notation

$P(x_i)$ to stand for $\mathbb{P}(X_i = x_i), x_i \in \mathcal{X}$, and similarly for a set $A \subseteq \{1, 2, \ldots, n\}$, we define $P(x_A) \triangleq \mathbb{P}(X_A = x_A), x_a \in \mathcal{X}^{|A|}$, where $X_A \triangleq \{X_i | i \in A\}$.

### 2.1 Graphical Models and Structure Learning

As mentioned before, an undirected graphical model corresponding to a probability distribution is specified by an undirected graph $G = (V, E)$, with each vertex $v_i \in V$ corresponding to a random variable $X_i$ which is a component of a $p$-dimensional random vector $X$ (for ease of notation, henceforth when we mention a node, we refer to the physical node in the graph, and the associated random variable. The exact meaning should be clear from the context). The edges $E \subseteq V \times V$ of a graphical model can be viewed as encoding the probability distribution of $X$ in several ways, all of which are equivalent under certain conditions. For the purposes of structure learning, an important interpretation is the *local Markov* property, stated below.

**Definition 1** ***(Local Markov)*** *Given $G(V, E)$, let $N(i) = \{j \in V | (i, j) \in E\}$ denote the neighborhood of node $i$. Then a random vector $X$ is said to obey the local Markov property with respect to the graph $G$ if for every $X_i \in V$, conditioned on the nodes in the neighborhood of $i$, the node $i$ is independent of the remaining nodes in the graph. Mathematically, this means that for any set $B \in V \setminus \{i\} \cup N(i)$, we have that $P(x_i | x_{N(i)}, x_B) = P(x_i | x_{N(i)})$ for all $(x_i, x_{N(i)}, x_B) \in \mathcal{X}^{1+|N(i)|+|B|}$. We henceforth write this as $X_i \overset{X_{N(i)}}{\perp\!\!\!\perp} X_{V \setminus \{i\} \cup N(i)}$.*

Finally, the structure learning problem is stated formally as follows: *given $n$ i.i.d. samples drawn from a random variable $X$ with MRF $G$, give a learning algorithm and associated conditions such that the hypothesis of the algorithm, $\hat{G}$, is equal to the true MRF $G$ with probability greater than $1 - \delta$.*

### 2.2 Factor Graphs

Every graphical model has a factor graph representation defined as follows.

**Definition 2** ***(Factor Graph)*** *Given a graphical model $G(V, E)$ its factor graph is a bipartite graph $G_f$ with vertex set $V \cup C$ where each vertex $c \in C$ corresponds to a maximal clique in $G$. For any $v \in V$ and $c \in C$, there is an edge $\{v, c\}$ in $G_f$ if and only if $v \in c$ in $G$.*

We have the following simple lemma relating the distance between two nodes $i, j \in V$ in the graphs $G$ and $G_f$.

**Lemma 1** *Given a graph $G$, let $G_f$ be its factor graph. Then for every $i, j \in V$ we have $d_f(i, j) = 2d(i, j)$ where $d$ and $d_f$ are the distances between $i$ and $j$ in $G$ and $G_f$ respectively.*

### 2.3 Conditional Entropy Tests

As we described before in the introduction, a comparison-test based method of structure learning is based on using a test function to compare candidate graphs. Although there are several different implementations, they are all based on the local Markov interpretation of

the graph. More specifically, most comparison-test algorithms try to learn the neighborhood of each individual node by comparing potential neighborhoods using a test function. Following the approach of Abbeel et. al.(Abbeel et al., 2006), we use conditional entropies as our test function for selecting nodes. In this section, we provide the necessary definitions, and also state some results from information theory that underlies our approach.

First we need to define a few quantities which we use throughout this paper. Given a discrete-valued random variable $Y$ taking values in a finite set $\mathcal{Y}$ such that $\mathbb{P}(Y = y) = p_y \geq 0 \, \forall \, y \in \mathcal{Y}$, and given $n$ i.i.d samples $\{Y^{(i)}\}_{i=1}^n$, the empirical probability mass function $\widehat{P}(y), y \in \mathcal{Y}$ is defined as,

$$\widehat{P}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y^{(i)}=y\}}, \, \forall \, y \in \mathcal{Y}.$$

The empirical entropy $\widehat{H}(Y)$ is defined as the entropy of the empirical distribution $\widehat{P}$.

Next, given two variables $Y_1, Y_2$, both taking values in $\mathcal{Y}$, we can extend this notation to define empirical conditional measures of the form

$$\widehat{P}(y_1|y_2) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{Y_1^{(i)}=y_1, Y_2^{(i)}=y_2\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Y_2^{(i)}=y_2\}}}, \, \forall \, (y_1, y_2) \in \mathcal{Y}^2.$$

Finally, for fixed $y_2 \in \mathcal{Y}$ we define empirical conditional entropy

$$\widehat{H}(Y_1|Y_2 = y_2) = -\sum_{y_1 \in \mathcal{Y}} \widehat{P}(y_1|y_2) \log \widehat{P}(y_1|y_2),$$

and using this we define,

$$\widehat{H}(Y_1|Y_2) = \sum_{y_2 \in \mathcal{Y}} \widehat{P}(y_2) \widehat{H}(Y_1|y_2)$$

Given samples, we use the empirical conditional entropies as given above as the proxy for the actual conditional entropy. Note also that we can define set based versions of all the above statements in a similar manner.

The use of conditional entropies as a test function is motivated by two reasons:

1. By the local Markov property, the conditional entropy for a node is minimized by sets which contain the true neighborhood, and hence (under some weak non-degeneracy conditions), the smallest cardinality set which minimizes the conditional entropy is the true neighborhood.

2. Entropy and measure are related in the sense that two probability measures on a set are close if their entropies are close and vice versa.

The first point is the main reason behind using conditional entropies as a test function, as it reduces the problem of finding a neighborhood to that of finding a set which minimizes an appropriate function, and also indicates a natural greedy sequential approach to selecting the neighbors. We encode this notion in the following proposition, which can be easily derived from the Data Processing Inequality, see (Cover and Thomas, 2006).

**Proposition 1** *For any node $i \in V$, we have that,*

$$H(X_i|X_{N(i)}) \leq H(X_i|X_A),$$

*for any set $A \subseteq V \setminus \{i\}$.*

The second point can be thought of as indicating that no information is lost if we use entropies instead of measures to learn the structure. This notion can be quantified in terms of the following proposition, which we get by combining Theorem 16.3.2 and Lemma 16.3.1 from (Cover and Thomas, 2006).

**Proposition 2** *Let $P$ and $Q$ be two probability mass functions in a finite set $\mathcal{X}$, with entropies $H(P)$ and $H(Q)$ respectively, and with total variational distance $||P - Q||_1$ given by:*

$$||P - Q||_1 = \sum_{x \in \mathcal{X}} |P(x) - Q(x)|.$$

*Then*

$$|H(P) - H(Q)| \leq -||P - Q||_1 \log \frac{||P - Q||_1}{|\mathcal{X}|}. \tag{1}$$

*Further, if the relative entropy between them is given by $D(P||Q)$, then*

$$D(P||Q) \geq \frac{1}{2 \log 2} ||P - Q||_1^2. \tag{2}$$

We use this proposition in several places in subsequent proofs. At a high level, (1) allows us to leverage results of convergence of empirical measures to the true measure to obtain similar guarantees on the empirical entropy, while (2) is used to convert entropy conditions to equivalent conditions on the measure (in particular, this allows us to state our non-degeneracy conditions directly in terms of the conditional entropy, instead of more complicated statements in terms of probability distributions usually found in literature (Bresler et al., 2008)).

## 3. The GreedyAlgorithm($\epsilon$) Structure Learning Algorithm

In this section, we present our greedy structure learning algorithm, which we henceforth refer to as
GreedyAlgorithm($\epsilon$). We also argue that it always has a low *worst-case* computation complexity, owing to its greedy nature. The challenge however is to find conditions that guarantee correctness, and this question is addressed in subsequent sections.

At a high level, our algorithm considers each node separately, and adds nodes to its neighborhood sequentially in a greedy manner. In particular, at each step we find the node that provides the highest reduction in conditional entropy when added to the existing set. We stop when this reduction is smaller than $\epsilon$.

More specifically, GreedyAlgorithm($\epsilon$) takes as input the $n$ samples and a single 'threshold' value $\epsilon$. Given any node $i$, the candidate neighborhood $\widehat{N}(i)$ of the node is initially set to $\phi$ and is learnt in a sequential manner. In the first stage, the node $j \neq i$ which minimizes the conditional entropy $H(X_i|X_j)$ is chosen as a candidate neighbor, and is added to $\widehat{N}(i)$

if conditioning on the node $j$ reduces the entropy by at-least $\epsilon/2$. In any subsequent stage, a candidate node $k \in V \setminus \widehat{N}(i)$ is chosen as one which minimizes $H(X_i|X_k, H_{\widehat{N}(i)})$, and is added if it reduces the conditional entropy by at-least $\epsilon/2$. At any stage when this condition is not satisfied, the algorithm outputs $\widehat{N}(i)$ and moves on to the next node.

GreedyAlgorithm($\epsilon$) for structure learning is formally presented in Algorithm 1.

**Algorithm 1** GreedyAlgorithm($\epsilon$)

1: **for** $i \in V$ **do**
2: complete $\leftarrow$ FALSE
3: $\widehat{N}(i) \leftarrow \Phi$
4: **while** !complete **do**
5: $j = \underset{k \in V \setminus \widehat{N}(i)}{\operatorname{argmin}} \widehat{H}(X_i \mid X_{\widehat{N}(i)}, X_k)$
6: **if** $\widehat{H}(X_i \mid X_{\widehat{N}(i)}, X_j) < \widehat{H}(X_i \mid X_{\widehat{N}(i)}) - \frac{\epsilon}{2}$ **then**
7: $\widehat{N}(i) \leftarrow \widehat{N}(i) \cup \{j\}$
8: **else**
9: complete $\leftarrow$ TRUE
10: **end if**
11: **end while**
12: **end for**

Since the algorithm is greedy, we can characterize its worst case computational complexity independent of its correctness guarantees.

**Proposition 3** *The running time of Algorithm 1 is $O(np^4)$ where $n$ is the number of samples and $p$ is the number of random variables.*

**Proof** The outer *for* loop is executed $O(p)$ times. For every iteration of the outer *for* loop, the *while* loop (lines 4-11) is run $O(p)$ times. In every iteration of the *while* loop, line 5 calculates the empirical entropy conditioned on each of the nodes in $\widehat{N}(i)$. Thus, in the worst case, the algorithm performs $O(p^3)$ comparison tests (empirical conditional entropy calculation from samples). Even assuming a naive implementation of a single comparison test that takes $O(np)$, the overall time taken by the algorithm is $O(np^4)$. ■

This shows that GreedyAlgorithm($\epsilon$) always has low computational complexity for any graph (and in particular, in Section 4, we show that for a large class of graphs, the algorithm has running time of $O(np^2)$). The tradeoff is however in correctness guarantees. The problem arises in the fact that unlike other comparison-test algorithms which are designed to ensure certain correctness guarantees, our algorithm is designed more from the point of view of simplicity and low computational costs. Therefore to derive theoretical guarantees for the algorithm, it is first important to understand the failure mechanism of the algorithm.

## 4. Sufficient Conditions for General Discrete Graphical Models

In this section, we provide guarantees for general discrete graphical models, under which GreedyAlgorithm($\epsilon$) recovers the graphical model structure exactly. First, using an exam-

ple, we build up intuition for the sufficient conditions, and define two key notions: non-degeneracy conditions and correlation decay. Our main result is presented in Section 4.2, wherein we give a sufficient condition for the correctness of the algorithm in general discrete graphical models.

### 4.1 Non-Degeneracy and Correlation Decay

Before analyzing the correctness of structure learning from samples, a simpler problem worth considering is one of algorithm consistency, i.e., does the algorithm succeed to identify the true graph *given the true conditional distributions* (or in other words, given an infinite number of samples). It turns out that the algorithm as presented in Algorithm 1 does not even possess this property, as is illustrated by the following counter-example

Let $V = \{0, 1, \cdots, D, D{+}1\}$, $X_i \in \{-1, 1\} \forall i \in V$ and $E = \{\{0, i\}, \{i, D+1\} \mid 1 \leq i \leq D\}$. Let $P(x_V) = \frac{1}{Z} \prod_{\{i,j\} \in E} e^{\theta x_i x_j}$, where $Z$ is a normalizing constant (this is the classical zero-field Ising model potential). The graph is shown in Fig. 1.

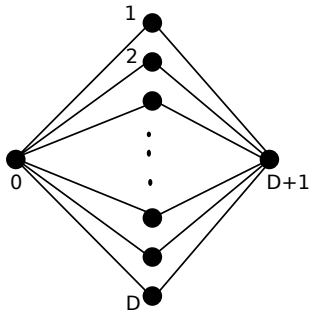



Figure 1: An example of adding spurious nodes: Execution of GreedyAlgorithm($\epsilon$) for node 0 adds node $D + 1$ in the first iteration, even though it is not a neighbor.

Suppose the actual entropies are given as input to Algorithm 1. It can be shown in this case that for a given $\theta$, there exists a $D_{\text{thresh}}$ such that if $D > D_{\text{thresh}}$, then the output of Algorithm 1 will select the edge $\{0, D + 1\}$ in the first iteration. This is easily understood because if $D$ is large, the distribution of node 0 is best accounted for by node $D + 1$, although it is not a neighbor. Thus, even with exact entropies, the algorithm will always include edge $(0, D + 1)$, although it does not exist in the graph.

The algorithm can however easily be shown to satisfy the following weaker consistency guarantee: given infinite samples, for any node in the graph, the algorithm will return a *super-neighborhood*, i.e., a superset of the neighborhood of $i$. This suggests a simple fix to obtain a consistent algorithm, as we can follow the greedy phase by a 'node-pruning' phase, wherein we test each node in the neighborhood of a node $i$ returned by the algorithm (to do this, we can compare the entropy of $i$ conditioned on the neighborhood with and without a node, and remove it if they are the same). However the problem is complicated by the presence of samples, as pruning a large super-neighborhood requires calculating estimates of entropy conditioned on a large number of nodes, and hence this drives up the sample complexity. In the rest of the paper, we avoid this problem by ignoring the pruning step,

and instead prove a stronger correctness guarantee: given any node $i$, the algorithm always picks a *correct* neighbor of $i$ as long as any one remains undiscovered. Towards this end, we first define two conditions which we require for the correctness of GreedyAlgorithm($\epsilon$).

**Assumption 1 (Non-degeneracy)** *Choose a node $i$. Let $N(i)$ be the set of its neighbors. Then $\exists \epsilon > 0$ such that $\forall\ A \subset N(i)$, $\forall\ j \in N(i) \setminus A$ and $\forall\ l \in N(j) \setminus \{i\}$, we have that*

$$H(X_i \mid X_A) - H(X_i \mid X_A, X_j) > \epsilon \text{ and} \tag{3}$$

$$H(X_i \mid X_A, X_l) - H(X_i \mid X_A, X_j, X_l) > \epsilon \tag{4}$$

Assumption 1 is illustrated in Fig. 2.

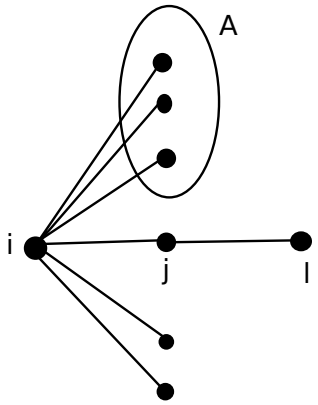



Figure 2: Non-degeneracy condition for node $i$: $(i)$ Entropy of $i$ conditioned on any sub-neighborhood $A$ reduces by at-least $\epsilon$ if any other neighbor $j$ is added to the conditioning set, $(ii)$ Entropy of $i$ conditioned on $A$ and a two hop neighbor $l$ reduces by at-least $\epsilon$ if the corresponding one hop neighbor $j$ is added to the conditioning set

**Assumption 2 (Correlation Decay)** *Choose a node $i$. Let $N^1(i)$ and $N^2(i)$ be the sets of its 1-hop and 2-hop neighbors respectively. Choose another set of nodes $B$. Let $d(i, B) = \min_{j \in B} d(i, j)$, where $d(i, j)$ denotes the distance between nodes $i$ and $j$. Then, we have that $\forall x_i, x_{N^1(i)}, x_{N^2(i)}, x_B$*

$$\left|P(x_i, x_{N^1(i)}, x_{N^2(i)} \mid x_B) - P(x_i, x_{N^1(i)}, x_{N^2(i)})\right| < f(d(i, B))$$

*where $f$ is a monotonic decreasing function.*

Assumption 1 (or a variant thereof) is a standard assumption for showing correctness of any structure learning algorithm, as it ensures that there is a *unique* minimal graphical model for the distribution from which the samples are generated. Although the way we state the assumption is tailored to our algorithm, it can be shown to be equivalent to similar assumptions in literature(Bresler et al., 2008). Informally speaking, Assumption 1 states that for node $i$, any 2-hop neighbor captures less information about node $i$ than the corresponding 1-hop neighbor. In the case of a Markov Chain, Assumption 1 reduces to a

weaker version of an $\epsilon-$Data Processing Inequality (i.e., DPI with an epsilon gap), and in a sense, Assumption 1 can be viewed as a generalized $\epsilon-$DPI for networks with cycles.

On the other hand, Assumption 2 along with large girth implies that the information a node $j$ has about node $i$ is 'almost Markov' along the shortest path between $i$ and $j$. This in conjunction with Assumption 1 implies that for any two nodes $i$ and $k$, the information about $i$ captured by $k$ is less than that captured by $j$ where $j$ is the neighbor of $i$ on the shortest path between $i$ and $k$.

### 4.2 Guarantees for the Recovery of a General Graphical Model

We now state our main theorem, wherein we give a sufficient condition for correctness of GreedyAlgorithm($\epsilon$) in a general graphical model.

The counter-example given in Section 4.1 suggests that the addition of spurious nodes to the neighborhood of $i$ is related to the existence of non-neighboring nodes of $i$ which somehow accumulate sufficient influence over it. The accumulation of influence is due to slow decay of influence on short paths (corresponding to a high $\theta$ in the example), and the effect of a large number of such paths (corresponding to high $D$). Correlation decay (Assumption 2) allows us to control the first. Intuitively, the second can be controlled if the neighborhood of $i$ is 'locally tree-like'. To quantify this notion, we define the girth of a graph Girth($G$) to be the length of the smallest cycle in the graph $G$. Now we have the following theorem.

**Theorem 2** *Consider a graphical model $G$ where the random variable corresponding to each node takes values in a set $\mathcal{X}$ and satisfies the following:*

- *Non-degeneracy (Assumption 1) with parameter $\epsilon$,*
- *Correlation decay (Assumption 2) with decay function $f(\cdot)$,*
- *Maximum degree $D$.*

*Define $h \triangleq h(\epsilon, D) \triangleq \frac{\epsilon^2 |\mathcal{X}|^{-2(D+1)^2}}{64}$ and suppose $f^{-1}(h)$ exists. Further suppose $G_f$ (the factor graph of $G$) obeys the following condition:*

$$Girth(G_f) \triangleq g_f > 4\left(f^{-1}(h) + 1\right). \tag{5}$$

*Then, given $\delta > 0$, GreedyAlgorithm($\epsilon$) recovers $G$ exactly with probability greater than $1-\delta$ with sample complexity $n = \xi\left(\epsilon^{-4} \log \frac{p}{\delta}\right)$, where $\xi$ is a constant independent of $p, \epsilon$ and $\delta$.*

The proof follows from the following two lemmas. Lemma 3 implies that if we had access to actual entropies, Algorithm 1 always recovers the neighborhood of a node exactly. Lemma 4 shows that with the number of samples $n$ as stated in Theorem 2, the empirical entropies are very close to the actual entropies with high probability and hence Algorithm 1 recovers the graphical model structure exactly with high probability even with empirical entropies.

**Lemma 3** *Consider a graphical model $G$ in which node $i$ satisfies Assumptions 1 and 2. Let the girth of $G_f$ be $g_f > 4\left(f^{-1}(h) + 1\right)$, where $h$ is as defined in Theorem 2. Then, $\forall\, A \subset N(i)$, $u \notin N(i)$, $\exists\, j \in N(i) \setminus A$ such that*

$$H(X_i \mid X_A, X_j) < H(X_i \mid X_A, X_u) - \frac{3\epsilon}{4} \tag{6}$$

**Proof** If $A$ separates $i$ and $u$ in $G_f$ it also does so in $G$. Then we have that $P(x_i|x_A, x_u) = P(x_i|x_A)$ and hence $H(X_i \mid X_A, X_u) = H(X_i \mid X_A)$. Then, the statement of the lemma follows from (3).

Now suppose $A$ does not separate $i$ and $u$ in $G_f$. Consider the shortest path between $i$ and $u$ in $G_f \setminus A$. Let $j \in N(i) \setminus A$ and $l \in N(j) \setminus \{i\}$ be on that shortest path. Assumption 1 implies that $H(X_i \mid X_A, X_l) - H(X_i \mid X_A, X_j, X_l) > \epsilon$. Now, choose $B \in V$ such that $A \cup B \cup \{j\}$ separates $i$ and $l$ in $G_f$ and $d_f(i, B) \geq \frac{g_f - 4}{2}$, where $g_f$ is the girth of $G_f$. Note that such a $B$ (possibly empty) exists since the girth of $G_f$ is $g_f$ and if a node in the separator is a factor node (i.e., not in $V$) then we can replace it by all its neighbors (in $V$). We then see using Lemma 1 that $d(i, B) \geq \frac{g_f - 4}{4}$. From Assumption 2, we know that

$$
\begin{aligned}
&|P(x_i, x_{N(i)\cup N^2(i)}) - P(x_i, x_{N(i)\cup N^2(i)} \mid x_B)| < f\left(\tfrac{g_f}{4} - 1\right) \\
&\Rightarrow \sum_{x_i, x_A, x_j} |P(x_i, x_A, x_j) - P(x_i, x_A, x_j \mid x_B)| < |\mathcal{X}|^{(D+1)^2} f\left(\frac{g_f}{4} - 1\right) \quad \forall\, x_B \\
&\Rightarrow H(X_i, X_A, X_j) - H(X_i, X_A, X_j \mid X_B) < -|\mathcal{X}|^{(D+1)^2} f\left(\tfrac{g_f}{4} - 1\right)\left(\log f\left(\tfrac{g_f}{4} - 1\right)\right) \triangleq \widehat{\epsilon} \\
&\Rightarrow (H(X_i \mid X_A, X_j) + H(X_A, X_j)) - (H(X_i \mid X_A, X_j, X_B) + H(X_A, X_j \mid X_B)) < \widehat{\epsilon} \\
&\Rightarrow H(X_i \mid X_A, X_j) - H(X_i \mid X_A, X_j, X_B) < \widehat{\epsilon},
\end{aligned}
$$

where the first implication follows from marginalizing irrelevant variables and the second implication follows from (1). Using this we have that,

$$
\begin{aligned}
H(X_i \mid X_A, X_j, X_l) &\geq H(X_i \mid X_A, X_j, X_l, X_B) \\
&= H(X_i \mid X_A, X_j, X_B) \text{ since } X_i \overset{X_A, X_j, X_B}{\perp\!\!\!\perp} X_l \\
&> H(X_i \mid X_A, X_j) - \widehat{\epsilon}
\end{aligned}
$$

Using a similar argument, we also have,

$$
H(X_i \mid X_A, X_l, X_u) > H(X_i \mid X_A, X_l) - \widehat{\epsilon}
$$

Combining the two inequalities, and using the fact that under the given conditions $\widehat{\epsilon} < \frac{\epsilon}{8}$, we get

$$
H(X_i \mid X_A, X_j) \leq H(X_i \mid X_A, X_u) - \frac{3\epsilon}{4}.
$$

■

**Lemma 4** *Consider a graphical model $G$ in which each node takes values in $\mathcal{X}$. Let the number of samples be*

$$
n > 2^{15} \epsilon^{-4} |\mathcal{X}|^{4(D+2)} \left((D+2)\log 2|\mathcal{X}| + 2\log\frac{p}{\delta}\right)
$$

*Let $\widehat{P}$ and $\widehat{H}$ denote the empirical probability and empirical entropy as defined in Section 2.3.*

*Then $\forall\, i \in G$, with probability greater than $1 - \frac{\delta}{p}$, we have that $\forall\, A \subseteq N(i),\ u \notin N(i)$*

$$
\left| H(X_i \mid X_A, X_u) - \widehat{H}(X_i \mid X_A, X_u) \right| < \frac{\epsilon}{8}
$$

**Proof** We use the fact that given sufficient samples, the empirical measure is close to the true measure uniformly in probability. Specifically, given any subset $A \subseteq V$ of nodes and any fixed $x_A \in \mathcal{X}^{|A|}$, we have by Azuma's inequality after $n$ samples,

$$\mathbb{P}\left[\left|P(x_A) - \widehat{P}(x_A)\right| > \gamma\right] < 2\exp(-2\gamma^2 n) < \frac{2\delta}{p^2(2|\mathcal{X}|)^{(D+2)}}.$$

where $\gamma = 2^{-8}\epsilon^2|\mathcal{X}|^{-2(D+2)}$. Let $V$ be the set of all vertices. Now, by union bound over every $A \subseteq N(i),\ u \in V$ and $x_i, x_A, x_u$, we have

$$\mathbb{P}\left[\left|P(x_i, x_A, x_u) - \widehat{P}(x_i, x_A, x_u)\right| > \gamma\right] < \frac{\delta}{p}.$$

(1) then implies

$$\mathbb{P}\left[\left|H(X_i \mid X_A, X_u) - \widehat{H}(X_i \mid X_A, X_u)\right| > \frac{\epsilon}{8}\right] < \frac{\delta}{p}.$$

giving us the required result. ■

Using Lemmas 3 and 4, we have the following : $\forall\ i \in G$, such that Assumptions 1 and 2 are satisfied, with probability greater than $1 - \frac{\delta}{p}$, we have that $\forall\ A \subseteq N(i),\ u \notin N(i),\ \exists\ j \in N(i) \setminus A$ such that

$$\widehat{H}(X_i \mid X_A, X_j) < \widehat{H}(X_i \mid X_A, X_u) - \frac{\epsilon}{2} \tag{7}$$

and $\forall\ i \in G$, such that Assumptions 1 and 2 are satisfied, $\forall A \subset N(i),\ j \in N(i) \setminus A$, we have that

$$\widehat{H}(X_i \mid X_A, X_j) < \widehat{H}(X_i \mid X_A) - \frac{\epsilon}{2} \tag{8}$$

**Proof** [Theorem 2] The proof is based on mathematical induction. The induction claim is as follows: just before entering an iteration of the WHILE loop, $\widehat{N}(i) \subset N(i)$. Clearly this is true at the start of the WHILE loop since $\widehat{N}(i) = \Phi$. Suppose it is true just after entering the $k^{\text{th}}$ iteration. If $\widehat{N}(i) = N(i)$ then clearly $\forall j \in V \setminus \widehat{N}(i),\ H(X_i \mid X_{\widehat{N}(i)}, X_j) = H(X_i \mid X_{\widehat{N}(i)})$. Since with probability greater than $1 - \frac{\delta}{p}$ we have that $\left|\widehat{H}(X_i \mid X_{\widehat{N}(i)}, X_j) - H(X_i \mid X_{\widehat{N}(i)}, X_j)\right| < \frac{\epsilon}{8}$ and $\left|\widehat{H}(X_i \mid X_{\widehat{N}(i)}) - H(X_i \mid X_{\widehat{N}(i)})\right| < \frac{\epsilon}{8}$, we also have that $\left|\widehat{H}(X_i \mid X_{\widehat{N}(i)}, X_j) - \widehat{H}(X_i \mid X_{\widehat{N}(i)})\right| < \frac{\epsilon}{4}$. So control exits the loop without changing $\widehat{N}(i)$. On the other hand, if $\exists j \in N(i) \setminus \widehat{N}(i)$ then from (8) we have that $\widehat{H}(X_i \mid X_{\widehat{N}(i)}) - \widehat{H}(X_i \mid X_{\widehat{N}(i)}, X_j) > \frac{\epsilon}{2}$. So, a node is chosen to be added to $\widehat{N}(i)$ and control does not exit the loop. Now suppose for contradiction that a node $u \notin N(i)$ is added to $\widehat{N}(i)$. Then we have that $\widehat{H}(X_i \mid X_{\widehat{N}(i)}, X_u) < \widehat{H}(X_i \mid X_{\widehat{N}(i)}, X_j)$. But this contradicts (7). Thus, a neighbor $j \in N(i) \setminus \widehat{N}(i)$ is picked in the iteration to be added to $\widehat{N}(i)$, proving that the neighborhood of $i$ is recovered exactly with probability greater than $1 - \frac{\delta}{p}$. Using union bound, it is easy to see that the neighborhood of each node (i.e., the graph structure) is recovered exactly with probability greater than $1 - \delta$. ■

**Remark 5** *The proof for Theorem 2 can also be used to provide node-wise guarantees, i.e., for every node satisfying Assumptions 1 and 2, if the number of samples is sufficiently large (in terms of its degree, and the length of the smallest cycle it is part of), its neighborhood will be recovered exactly with high probability.*

**Remark 6** *Any decreasing correlation-decay function $f$ suffices for Theorem 2. However, the faster the correlation decay, the smaller the girth in the sufficient condition for Theorem 2 needs to be.*

And finally we have a corollary for the computational complexity of GreedyAlgorithm($\epsilon$).

**Corollary 1** *The expected run time of Algorithm 1 is $O\left(\delta np^4 + (1-\delta)D^2np^2\right)$. Further, if $\delta$ is chosen to be $O(p^{-2})$, the sample complexity $n$ is $O(\log p)$ and the expected run time of Algorithm 1 is $O(D^2p^2\log p)$.*

**Proof** For the second part, note that with probability greater than $1-\delta$, the algorithm recovers the correct graph structure exactly. In this case, the number of iterations of the *while* loop is bounded by $D$ for each node . The time taken to compute any conditional entropy is bounded by $O(nD)$. Hence the total run time is $O(D^2np^2)$. Using the previous worst case bound on the running time, we obtain the result. ■

## 5. Guarantees for the Recovery of an Ising Graphical Model

In this section, we show how Theorem 2 can be used to efficiently learn Ising graphical models satisfying certain conditions. The zero field Ising model is a pairwise, symmetric, binary graphical model which is widely used in statistical physics to model the alignment of magnetic spins in a magnetic field (Brush, 1967). It is defined as follows:

**Definition 3** *A set of random variables $\{X_v \mid v \in V\}$ are said to be distributed according to a zero field Ising model if*

1. $X_v \in \{-1, 1\}\ \forall v \in V$ *and*

2. $P(x_V) = \frac{1}{Z}\prod_{i,j\in V}\exp(\theta_{ij}x_ix_j)$

*where $Z$ is a normalizing constant. The graphical model of such a set of random variables is given by $G(V, E)$ where $E = \{\{i, j\} \mid \theta_{ij} \neq 0\}$.*

It is easy to verify that this satisfies the local Markov property. Another very useful property of zero-field Ising models is that they are symmetric with respect to $-1$ and 1. Formally, if $P$ is the probability distribution function over a set of zero-field Ising distributed random variables, then, $P(x_V) = P(-x_V)$.

The main contribution of this section is in the form of the following theorem, which translates the sufficient conditions from Section 4 to equivalent conditions for an Ising model.

**Theorem 7** *Consider a zero-field Ising model on a graph $G$ with maximum degree $D$. Let the edge parameters $\theta_{ij}$ be bounded in the absolute value by $0 < \beta < |\theta_{ij}| < \frac{\log 2}{2D}$. Let $\epsilon \triangleq 2^{-10} \sinh^2(2\beta)$. If the girth of the graph satisfies $g > \frac{2^{15}}{\log 2} \left\{ D^2 \log 2 - \log(\sinh 2\beta) \right\}$ then with samples $n = \xi \epsilon^{-4} \log \frac{p}{\delta}$ (where $\xi$ is a constant independent of $\epsilon, \delta, p$), GreedyAlgorithm($\epsilon$) outputs the exact structure of $G$ with probability greater than $1 - \delta$.*

The proof of this theorem consists of showing that an Ising graphical model satisfies Assumptions 1 and 2 if the graph has large girth and the parameters on the edges satisfy certain conditions. It also uses the fact that the girth $g_f$ of $G_f$ is at least $2g$. In Section 5.1, we show that under certain conditions, an Ising model has an almost exponential correlation decay. Then in Section 5.2, we use the correlation decay of Ising models to show that under some further conditions, they also satisfy Assumption 1 for non-degeneracy. Combining the two, we get the above sufficient conditions for GreedyAlgorithm($\epsilon$) to learn the structure of an Ising graphical model with high probability.

### 5.1 Correlation Decay in Ising Models

We will start by proving the validity of Assumption 2 in the form of the following proposition.

**Proposition 4** *Consider a zero-field Ising model on a graph $G$ with maximum degree $D$ and girth $g$. Let the edge parameters $\theta_{ij}$ be bounded in the absolute value by $|\theta_{ij}| < \frac{\log 2}{2D}$. Then, for any node $i$, its neighbors $N^1(i)$, its 2-hop neighbors $N^2(i)$ and a set of nodes $A$, we have*

$$\left| P(x_i, x_{N^1(i)}, x_{N^2(i)} \mid x_A) - P(x_i, x_{N^1(i)}, x_{N^2(i)}) \right| < c \exp\left( -\frac{\log 2}{3} \min\left( d(i, A), \frac{g}{2} - 1 \right) \right)$$

*$\forall\, x_i, x_{N^1(i)}, x_{N^2(i)}$ and $x_A$ (where $c$ is a constant independent of $i$ and $A$).*

The outline of the proof of Proposition 4 is as follows. First, we show that if a subset of nodes is conditioned on a Markov blanket (i.e., on another subset of nodes which separates them from the remaining graph), then their potentials remain the same. For this we have the following lemma.

**Lemma 8** *Consider a graphical model $G(V, E)$ and the corresponding factorizable probability distribution function $P$. Let $A, B$ and $C$ be a partition of $V$ and $B$ separate $A$ and $C$ in $G$. Let $\tilde{G}(A \cup B, \tilde{E})$ be the induced subgraph of $G$ on $A \cup B$, with the same edge potentials as $G$ on all its edges and $\tilde{P}$ be the corresponding probability distribution function. Then, we have that $P(x_D \mid x_B) = \tilde{P}(x_D \mid x_B)\ \forall\, x_D, x_B$ where $D \subseteq A$.*

Now, for any node $i$, the induced subgraph on all nodes which are at distance less than $\frac{g}{2} - 1$ is a tree. Thus we can concentrate on proving correlation decay for a tree Ising model. We do this through the following steps:

1. Without loss of generality, the tree Ising model can be assumed to have all positive edge parameters

2. The worst case configuration for the conditional probability of the root node is when all the leaf nodes are set to the same value and all the edge parameters are set to the maximum possible value

3. For this scenario, correlation decays exponentially

The following three lemmas encode these three steps. For proofs, refer the Appendix.

**Lemma 9** *Consider a tree Ising graphical model $T$. Let the corresponding probability distribution be $P$. Replace all the edge parameters on this graphical model by their absolute values. Let the corresponding probability distribution after this change be $\tilde{P}$. Then, there exists a set of bijections*
*$\{M_v : \{-1,1\} \to \{-1,1\} \mid v \in V \setminus \{r\}\}$ where $V$ is the set of vertices and $r$ is the root node such that, $\forall x_r, x_{V\setminus r}$ we have that $P(x_r, x_{V\setminus r}) = \tilde{P}(x_r, M_v(x_v), v \in V \setminus r)$.*

**Lemma 10** *For a tree Ising graphical model $T$ with root $r$ and set of leaves $L$, we have*

$$(x_r = 1, x_L = 1) \in \arg\max_{x_r, x_L} |P(x_r \mid x_L) - P(x_r)|$$

And finally we have the following lemma.

**Lemma 11** *In a tree Ising model, suppose $|\theta_{ij}| < \gamma < \frac{\log 2}{2D}$ where $D$ is the maximum degree of the graph. Then we have exponential correlation decay between the root node $r$, its neighbors $N^1(r)$, its 2-hop neighbors $N^2(r)$ and the set of leaves $L$ i.e.,*

$$\left|P(x_r, x_{N^1(r)}, x_{N^2(r)} \mid x_L) - P(x_r, x_{N^1(r)}, x_{N^2(r)})\right| < c \exp(-\frac{\log 2}{3} d(r, L))$$

*where $c$ is a constant independent of the nodes considered.*

### 5.2 Non-degeneracy in Ising Models with Correlation Decay

Now using the results from the previous section, we turn our attention to the question of correlation decay. In particular, we have the following lemma which says that if an Ising graphical model has almost exponential correlation decay and its edge parameters satisfy certain conditions, then it also satisfies Assumption 1. For the proof, refer the Appendix.

**Lemma 12** *Consider an Ising graphical model with edge parameters $\theta_{ij}$ bounded in the absolute value by $0 < \beta < |\theta_{ij}| < \gamma$, max degree $D$, and having correlation decay as follows*

$$\left|P(x_i, x_{N^1(i)}, x_{N^2(i)}) - P(x_i, x_{N^1(i)}, x_{N^2(i)} | x_B)\right| < c \exp\left(-\alpha \min\left(d(i,B), \frac{g-2}{2}\right)\right)$$

$\forall\ i, B, x_i, x_{N^1(i)}, x_{N^2(i)}$. *If the girth $g > 2 + \frac{2}{\alpha}\Big\{(2D+11)\log 2 + \log c + \log\left(1 + 2^D e^{2\gamma}\right) + 2\gamma(D+3) - \log|\sinh 2\beta|\Big\}$, then this graphical model satisfies Assumption 1 with $\epsilon = 2^{-7} e^{-6\gamma D} \sinh^2(2\beta)$.*

Finally, the proof of Theorem 7 follows directly by combining Theorem 2, Proposition 4 and Lemma 12. For complete details, refer the Appendix.

## 6. Simulations

In this section, we present the results of numerical experiments evaluating the performance of our algorithm. There are two important points to be noted here. The first is that to satisfy the conditions so that our theoretical guarantees are applicable, the graph should have a large girth. However, to demonstrate the fact that our algorithm is practical, we evaluate our algorithm on graphs with much smaller girth than what is required for our theoretical guarantees to hold. The second is that even when we satisfy the conditions for our theoretical guarantees to be applicable, we are confronted with the question of choosing $\epsilon$, which is an input to our algorithm. The nice behavior of our algorithm with respect to $\epsilon$ provides a partial solution to this problem by allowing us to choose a typical $\epsilon$ for the experiments. However, this also motivates the question of how to choose the value of $\epsilon$ experimentally, which will be interesting to look at in future work.

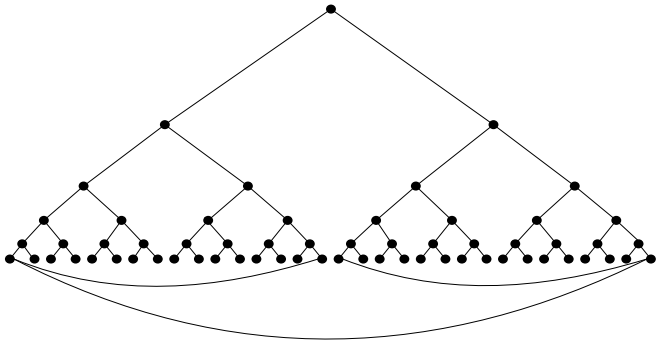
In the first experiment, we consider an Ising model on a binary tree of depth 5 with a few additional edges between the leaves. The graph is shown in Fig. 3(a). As remarked earlier, this graph does not satisfy the conditions (on girth) for our theoretical guarantees to be applicable. However, our algorithm seems to perform very well in learning this graphical model. This is not surprising because the graph has a structure favourable to our algorithm (i.e., large girth and moderate edge parameters, though they do not meet the conditions for our guarantees to hold). Fig. 3(b) presents the plots of probability of success versus number of samples of our algorithm for various values of $\epsilon$. Here, success is defined as exact recovery of the graph structure. There seems to exist a threshold value of $\epsilon$, call it $\bar{\epsilon}$ such that if $\epsilon > \bar{\epsilon}$ then the probability of success is very small and if $\epsilon < \bar{\epsilon}$, probability of success goes to 1 as the number of samples increases. This would suggest that the graph under consideration in fact satisfies Assumption 1 with $\bar{\epsilon}$. Fig. 4 presents the results of our algorithm (using a typical value of $\epsilon$) comparing it to the algorithm in (Ravikumar et al., 2010), which we will henceforth refer to as RWL.

In the second experiment, we evaluate our algorithm on grids of various sizes. Fig. 5 compares the sample complexity and computational complexity of our algorithm to RWL. From the figure, it is clear that our algorithm has higher sample complexity but lower computational complexity compared to RWL.

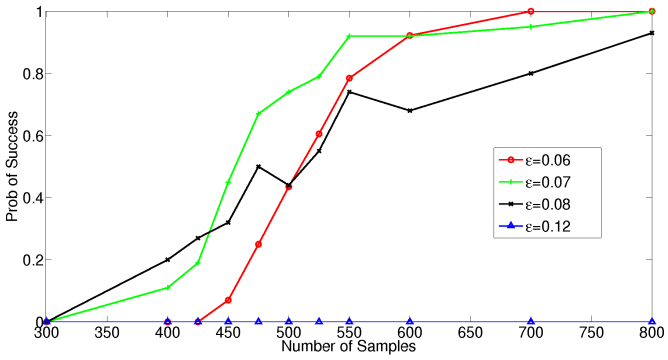
Finally, we present an application of our algorithm to model senator interaction graph using the senate voting records, following (Banerjee et al., 2008). A *Yea* vote is treated as a 1 where as a *Nay* vote or *absentee* vote is treated as $-1$. To avoid bias, we only consider senators who have voted in a fraction of atleast 0.75 of all the bills during the years 2009 and 2010. The output graph is presented in Fig. 6.

## 7. Discussion

We developed a simple greedy algorithm for Markov structure learning. The algorithm is simple to implement and has low computational complexity. We then showed that under some non-degeneracy, correlation decay, maximum degree and girth assumptions on the MRF, our algorithm recovers the correct graph structure with $O(\epsilon^{-4} \log \frac{p}{\delta})$ samples. We then specialize our conditions to prove a self-contained result for the most popular discrete graphical model - the Ising model.



(a)



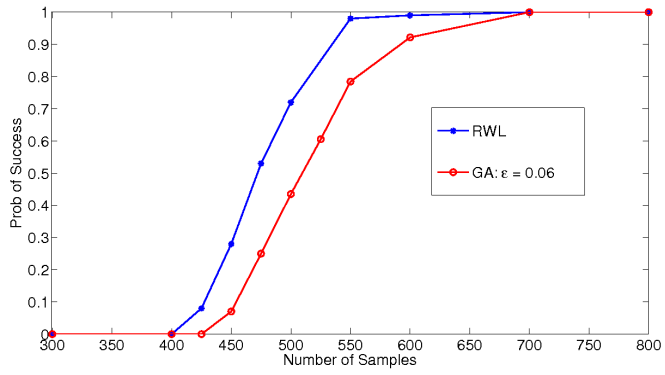

(b)

Figure 3: (a) The graph chosen for our experiments, binary tree with a few additional edges. (b) Results of our algorithm for various values of $\epsilon$. The edge parameters ($\theta_{ij}$) are all chosen to be equal to 0.5. Success is defined as exact recovery of the structure. The probability of success on the y-axis is calculated by averaging over 100 runs. For a large value of $\epsilon$, the probability of success of our algorithm is equal to 0. However, for smaller values of $\epsilon$, the probability of success goes to 1 as the number of samples increases.

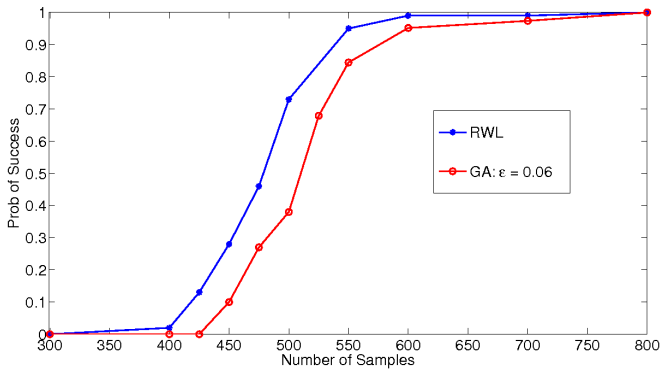
The success of our algorithm can be further improved by post-processing via *pruning*. In particular, as mentioned, the neighborhood of a node as estimated by our algorithm always includes the true neighborhood – but it may also include spurious nodes. The latter can be then identified by checking each node of the estimated neighborhood, to see if it actually provides a reduction in conditional entropy over and above all the other nodes. Analysis of the improvement achieved by such a procedure is more challenging, but it may be likely that doing so will reveal an algorithm that can handle much larger degrees and smaller girths.

## Acknowledgments

This work was partially supported by ARO grant W911NF-10-1-0360. We wish to thank Jason K. Johnson for letting us use his graph drawing code for the senator graph.




(a)




(b)

Figure 4: (a) The edge parameters are all chosen to be equal to 0.5. (b) The edge parameters are chosen uniformly at random from $\{-0.5, 0.5\}$.
GA refers to our algorithm. The probability of success on the y-axis is calculated by averaging over 100 runs. In both the cases, the sample complexity of our algorithm is slightly higher than that of RWL. However, our algorithm is more general (i.e., not specialized for an Ising model) and has lower computational complexity than RWL.

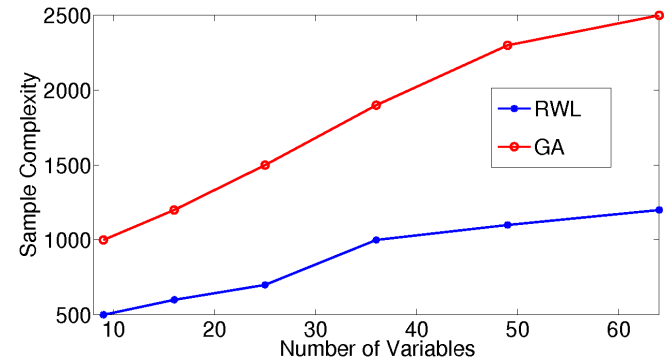
## References

P. Abbeel, D. Koller, and A. Y. Ng. Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.*, 7:1743–1788, 2006. ISSN 1532-4435.

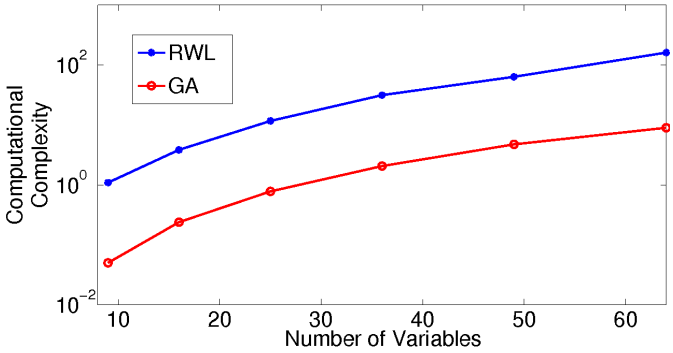
A. Anandkumar and V. Y. F. Tan. High-Dimensional Gaussian Graphical Model Selection: Tractable Graph Families. *Preprint*, June 2011a.

A. Anandkumar and V. Y. F. Tan. High-Dimensional Structure Learning of Ising Models : Tractable Graph Families. *Preprint*, June 2011b.

O Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9: 485–516, 2008. ISSN 1532-4435.




(a)




(b)

Figure 5: Plots of (a) sample complexity and (b) computational complexity for various grid sizes. Edge parameters are all chosen to be equal to 0.5. X-axis represents the number of variables (9 for a $3 \times 3$ grid, 16 for a $4 \times 4$ grid and so on). In (a), Y-axis represents the sample complexity which is taken to be the minimum number of samples required to obtain a probability of success of 0.95. In (b), Y-axis is in logarithmic scale and represents the time taken in seconds for a single run using the number of samples from (a). All the above quantities are calculated by averaging over 50 runs.

Jose Bento and Andrea Montanari. Which graphical models are difficult to learn?, 2009. http://arxiv.org/abs/0910.5761.

Andrej Bogdanov, Elchanan Mossel, and Salil P. Vadhan. The complexity of distinguishing markov random fields. In *APPROX-RANDOM*, pages 331–342, 2008.

G. Bresler, E. Mossel, and A. Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. In *APPROX '08 / RANDOM '08*, pages 343–356, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85362-6. doi: http://dx.doi.org/10.1007/978-3-540-85363-3_28.

Stephen G. Brush. History of the Lenz-Ising model. *Rev. Mod. Phys.*, 39(4):883–893, Oct 1967. doi: 10.1103/RevModPhys.39.883.

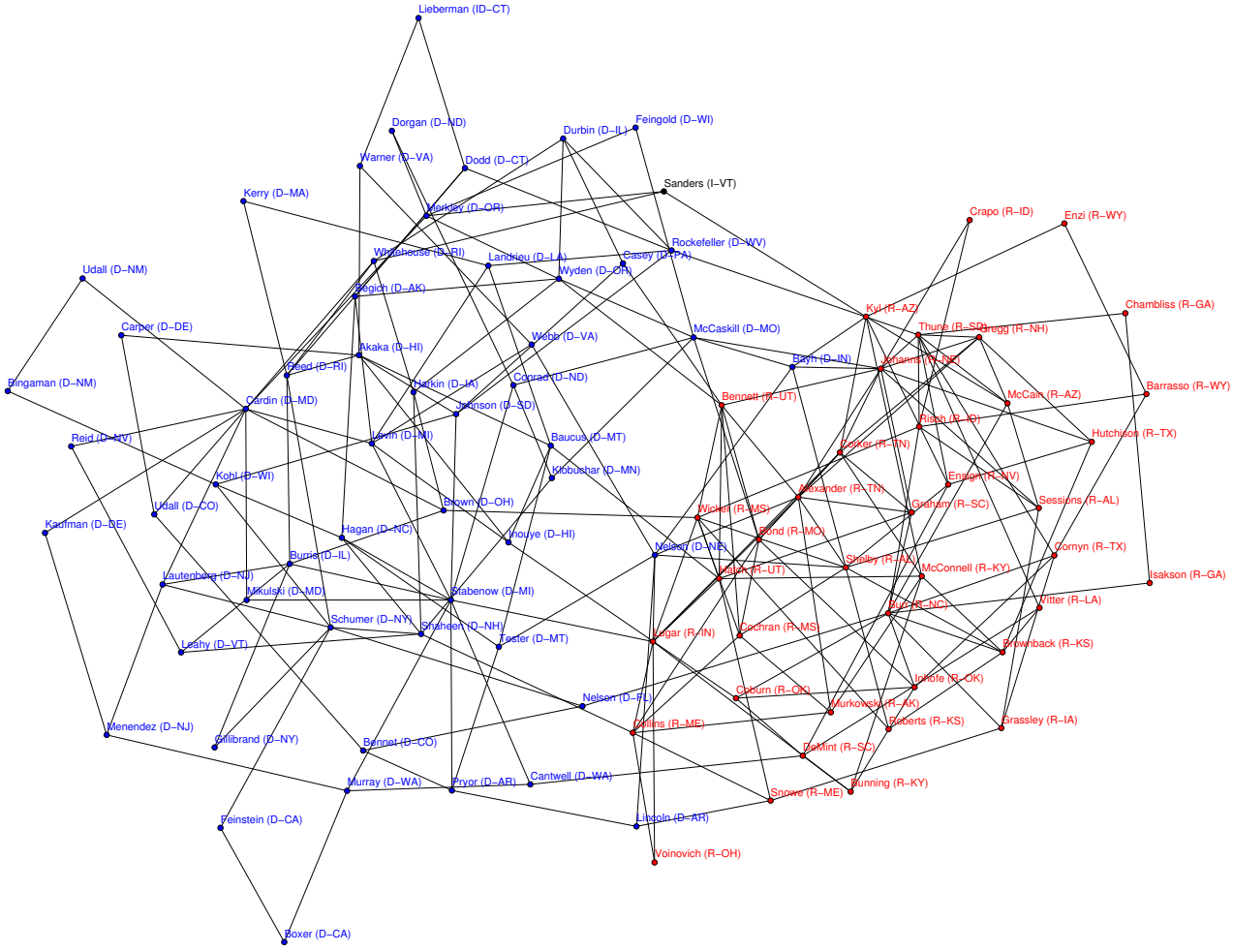



Figure 6: Blue nodes represent democrats, red nodes represent republicans and black node represents an independent. We use a value of 0.05 for $\epsilon$ in the algorithm. We can make some preliminary observations from the graph. Most of the democrats are connected to other democrats and most of the republicans are connected to other republicans (in particular, the number of edges between democrats and republicans is approximately 0.1 fraction of the total number of edges). The senate minority leader, McConnell is well connected to other republicans where as the senate majority leader, Reid is not well connected to other democrats. Sanders and Lieberman, both of who caucus with democrats have more edges to democrats than to republicans. We use the graph drawing algorithm of Kamada and Kawai to render the graph (Kamada and Kawai, 1989).

C. I. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory, 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. ISBN 0471241954.

T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Inf. Proc. Letters*, 31(12):7–15, 1989.

P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai. Greedy learning of markov network structure. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1295 –1302, Sept. 29 - Oct. 1 2010.

P. Ravikumar, M. W. Wainwright, and J. D. Lafferty. High-dimensional graphical model selection using $l_1$-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

Narayana P. Santhanam and Martin J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions, 2009. http://arxiv.org/abs/0905.2639.

N. Srebro. Maximum likelihood bounded tree-width markov networks. *Artificial Intelligence*, 143(1):123 – 138, 2003.

Vincent Y. F. Tan, Animashree Anandkumar, and Alan S. Willsky. Learning high-dimensional markov forest distributions: Analysis of error rates, 2010. http://arxiv.org/abs/1005.0766.

Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008. ISBN 1601981848, 9781601981844.

## Appendix

We will first prove the lemmas required for proving Proposition 4

**Proof** [Lemma 9] The proof is by construction. For each node $v \in V$, let $M_v(x_v) = \eta_v x_v$. For the root node, let $\eta_r \triangleq 1$. For any other node $v$, let $u$ be the parent of $v$ in the rooted tree with root $r$. Define $\eta_v \triangleq \frac{\theta_{uv}}{|\theta_{uv}|}\eta_u$. Let $\Phi$ and $\widetilde{\Phi}$ be the potential functions corresponding to $P$ and $\widetilde{P}$ respectively. Then,

$$\begin{aligned}
\Phi(x_V) &= \prod_{uv\in T} \exp\left(\theta_{uv}x_u x_v\right) \\
&= \prod_{uv\in T} \exp\left(|\theta_{uv}|\,\frac{\theta_{uv}}{|\theta_{uv}|}\eta_u^2 x_u x_v\right) \\
&= \prod_{uv\in T} \exp\left(|\theta_{uv}|\,\eta_u\eta_v x_u x_v\right) \\
&= \prod_{uv\in T} \exp\left(|\theta_{uv}|\,M_u(x_u)M_v(x_v)\right) \\
&= \widetilde{\Phi}(x_r, M_v(x_v), v \in V\setminus r)
\end{aligned}$$

Since the potential functions are preserved by the bijections, so are the probabilities. ■

We will first prove the following lemma which will help us in proving Lemma 10.

**Lemma 13** *Consider a tree Ising graphical model $T$ with root $r$, set of leaves $L$ and all positive edge parameters. Let $P$ be its probability distribution. Then, the quantity $P(X_r = 1 \mid X_L = x_L)$ is monotonically increasing in $x_l$, $\forall\, l \in L$. Moreover, $P(X_r = 1 \mid X_L = 1)$ is monotonically increasing in $\theta_{ij}$ $\forall\, \{i,j\} \in T$.*

**Proof** For simplicity of notation, we define $f(x_L) \triangleq P(X_r = 1 \mid X_L = x_L)$. Let us prove the above statement by induction on the depth of the tree. For a tree of depth 1, we have that

$$\begin{aligned} f(x_L) &= \frac{\prod_{l \in L} \exp(\theta_{rl} x_l)}{\prod_{l \in L} \exp(\theta_{rl} x_l) + \prod_{l \in L} \exp(-\theta_{rl} x_l)} \\ &= \frac{\prod_{l \in L, l \neq \widetilde{l}} \exp(\theta_{rl} x_l)}{\prod_{l \in L, l \neq \widetilde{l}} \exp(\theta_{rl} x_l) + \exp(-2\theta_{r\widetilde{l}} x_{\widetilde{l}}) \prod_{l \in L, l \neq \widetilde{l}} \exp(-\theta_{rl} x_l)} \end{aligned}$$

Since $\theta_{r\widetilde{l}} > 0$, $f(x_L)$ increases when $x_{\widetilde{l}}$ is changed from $-1$ to 1.

Now, suppose the statement is true for all trees of depth upto $k$. Consider a tree of depth $k+1$, with root $r$. Let $N(r)$ be the set of children of $r$. For every $c \in N(r)$, let $T_c$ be the subtree rooted at $c$ with the same edge parameters as in $T$ and $L_c$ be the leaves of $T_c$. Let $P_c$ be the probability measure corresponding to $T_c$ and $f_c(x_{L_c}) \triangleq P_c(x_c = 1 \mid x_{L_c})$. Then, the conditional probability of the root node can be written as

$$f(x_L) = \frac{\prod_{c \in N(r)} \left(\exp(\theta_{rc}) f_c(x_{L_c}) + \exp(-\theta_{rc}) \left(1 - f_c(x_{L_c})\right)\right)}{B} \tag{9}$$

where

$$\begin{aligned} B = &\prod_{c \in N(r)} \left(\exp(\theta_{rc}) f_c(x_{L_c}) + \exp(-\theta_{rc}) \left(1 - f_c(x_{L_c})\right)\right) + \\ &\prod_{c \in N(r)} \left(\exp(-\theta_{rc}) f_c(x_{L_c}) + \exp(\theta_{rc}) \left(1 - f_c(x_{L_c})\right)\right) \end{aligned}$$

(9) can now be manipulated to obtain (10).

$$f(x_L) = \frac{K_1}{K_1 + K_2 \frac{g_{\widetilde{c}}(x_{\widetilde{c}}) + \exp(2\theta_{r\widetilde{c}})}{g_{\widetilde{c}}(x_{\widetilde{c}}) \exp(2\theta_{r\widetilde{c}}) + 1}} \tag{10}$$

where $g_{\widetilde{c}}(x_{\widetilde{c}}) = \frac{f_{\widetilde{c}}(x_{L_{\widetilde{c}}})}{1 - f_{\widetilde{c}}(x_{L_{\widetilde{c}}})}$, and $K_1$ and $K_2 > 0$ are independent of $x_{L_{\widetilde{c}}}$ and $\theta_{r\widetilde{c}}$. Since $K_2 > 0$ and $\theta_{r\widetilde{c}} > 0$, $f(x_L)$ increases if $f_{\widetilde{c}}(x_{L_{\widetilde{c}}})$ increases. So, for any leaf node, if its value changes from $-1$ to 1, the corresponding $f_{\widetilde{c}}(x_{L_{\widetilde{c}}})$ increases and hence $f(x_L)$ increases, proving the induction claim.

Using the same induction argument as above and noting that $f(x_L = 1) > \frac{1}{2}$, it can be seen that $f(x_L = 1)$ is monotonically increasing in $\theta_{ij}\ \forall \{i,j\} \in T$. ■

**Proof** [Lemma 10] We know that $P(x_r) = \frac{1}{2}$ for $x_r = \pm 1$. Clearly any $x_L$ that maximizes $|P(x_r \mid x_L) - P(x_r)|$ should either minimize or maximize $P(x_r \mid x_L)$. Note also that there is a one-one correspondence between such configurations (i.e., for every maximizing configuration, there exists a minimizing configuration such that both of them maximize $|P(x_r \mid x_L) - P(x_r)|$). From Lemma 13, we know that $x_L = 1$ maximizes $P(x_r = 1 \mid x_L)$ and by symmetry this should be the same as $P(x_r = -1 \mid x_L = -1)$ and equal $\max_{x_L} P(x_r = -1 \mid x_L)$. So, we can conclude that $|P(x_r \mid x_L) - P(x_r)|$ is maximized by $(x_r = 1, x_L = 1)$. ■

**Lemma 14** *Consider a tree Ising model $T$ with root node $r$, set of leaves $L$ and maximum degree $D$. Let $P$ be its probability measure. Suppose the absolute values of the edge parameters are bounded by $|\theta_{ij}| < \frac{\log 2}{2D}\ \forall\ \{i,j\} \in T$. Then, we have that $|P(x_r \mid x_L) - P(x_r)| < \exp(-\frac{\log 2}{3} d(r, L))\ \forall x_r, x_L$.*

**Proof** Using Lemmas 9, 10 and 13, we can assume without loss of generality that the parameters $\theta_{ij}$ on all the edges are positive and equal to $\frac{\log 2}{2D}$ (which is the maximum possible value), consider a complete D-ary tree and concentrate on $|P(X_r = 1 \mid X_L = 1) - P(X_r = 1)|$. For simplicity of notation, let $\theta \triangleq \frac{\log 2}{2D}$. For a tree of depth $d$, let $a(d) \triangleq P(X_r = 1 \mid X_L = 1)$. We have that

$$a(d+1) = \frac{\left(\exp(\theta)a(d) + \exp(-\theta)\left(1 - a(d)\right)\right)^D}{\left(\exp(\theta)a(d) + \exp(-\theta)\left(1 - a(d)\right)\right)^D + \left(\exp(-\theta)a(d) + \exp(\theta)\left(1 - a(d)\right)\right)^D}$$

Using some algebraic manipulations and substituting the value of $\theta$, we obtain

$$\left|a(d+1) - \frac{1}{2}\right| < \exp\left(-\frac{\log 2}{3}\right)\left|a(d) - \frac{1}{2}\right|$$

and the result follows. ■

We need the following lemma to prove Lemma 11.

**Lemma 15** *Consider a tree Ising model $T$, with root node $r$, set of leaves $L$ and maximum degree $D$. Let $P$ be its probability measure. Suppose the absolute values of the edge parameters are bounded by $|\theta_{ij}| < \frac{\log 2}{2D}\ \forall\ \{i,j\} \in T$. Then, $\forall c$ such that $c$ is a child of $r$, we have that $|P(x_c \mid x_r, x_L) - P(x_c \mid x_r)| < 4\exp(-\frac{\log 2}{3} d(r, L))\ \forall x_r, x_j, x_L$.*

**Proof** Using Lemma 9 we can assume without loss of generality that the parameters $\theta_{ij}$ on all the edges are positive. $(x_c, x_r)$ can take values $(\pm 1, \pm 1)$. For each of those values, the value of $x_L$ that maximizes $|P(x_c \mid x_r, x_L) -$
$P(x_c \mid x_r)|$ either maximizes or minimizes $P(x_c \mid x_r, x_L)$. Noting from (a slight extension to) Lemma 13 that $P(x_c \mid x_r, x_L)$ is monotonic in $x_L$, it suffices to consider the eight possibilities $|P(X_c = \pm 1 \mid X_r = \pm 1, X_L = \pm 1) -$
$P(X_c = \pm 1 \mid X_r = \pm 1)|$. We show how to calculate the above value for $x_c = 1, x_r = 1, x_L =$

1. Interested readers can check that the conclusions below apply to all the other cases as well. Using Lemma 13, we can assume that the parameters $\theta_{ij}$ on all the edges except the edge $\{r, c\}$ are equal to $\frac{\log 2}{2D}$ and consider a complete D-ary tree. Let $\theta \triangleq \theta_{rc}$. We know that $P(X_c = 1 \mid X_r = 1) = \frac{\exp(\theta)}{\exp(\theta)+\exp(-\theta)}$. Let $d$ be the depth of the tree and $b(d) \triangleq P(X_c = 1 \mid X_r = 1, X_L = 1)$. We have $b(d) = \frac{\exp(\theta)a(d-1)}{\exp(\theta)a(d-1)+\exp(-\theta)(1-a(d-1))}$ where $a(d)$ is as defined in Lemma 14. Using some algebraic manipulations, it can be shown that $\left|b(d) - \frac{\exp(\theta)}{\exp(\theta)+\exp(-\theta)}\right| < 2\left|a(d-1) - \frac{1}{2}\right|$. Using Lemma 14 finishes the proof. ■

**Proof** [Lemma 11] Using Lemma 15, we have

$$\begin{aligned}
&\left|P(x_r, x_{N^1(r)}, x_{N^2(r)} \mid x_L) - P(x_r, x_{N^1(r)}, x_{N^2(r)})\right| \\
&= \Bigg| P(x_r \mid x_L) \prod_{j \in N^1(r)} P(x_j \mid x_r, x_L) \prod_{k \in N^2(r)} P(x_k \mid x_j, x_L) \\
&\quad - P(x_r) \prod_{j \in N^1(r)} P(x_j \mid x_r) \prod_{k \in N^2(r)} P(x_k \mid x_j) \Bigg| \\
&< 2^{D^2+3} \exp\left(-\tfrac{\log 2}{3}\left(d(r, L) - 1\right)\right) \\
&= c \exp\left(-\tfrac{\log 2}{3} d(r, L)\right)
\end{aligned}$$

proving the result. ■

**Proof** [Proposition 4] Let $I \triangleq \{i\} \cup N^1(i) \cup N^2(i)$. Let $B$ be a set that separates $I$ and $A$ such that $d(I, B) = \min(d(i, A), \frac{g}{2} - 1)$. Let $J$ be the component of nodes containing $I$ when the graph is separated by $B$. We know that the induced subgraph on $J \cup B$ is a tree. Applying Lemma 11 on this tree and using Lemma 8, we obtain $|P(x_I \mid x_B) - P(x_I \mid \widetilde{x}_B)| < 2c \exp(-\frac{\log 2}{3} d(I, B)) \ \forall x_I, x_B, \widetilde{x}_B$. Since $P(x_I)$ is a weighted average of $P(x_I \mid x_B)$ for various $x_B$, we have

$$|P(x_I \mid x_B) - P(x_I)| < 2c \exp(-\frac{\log 2}{3} d(I, B)) \ \forall x_I, x_B$$

The result then follows since $P(x_I \mid x_A)$ is a weighted average of $P(x_I \mid x_B)$. ■

**Proof** [Lemma 12] Let the graphical model be denoted by $G(V, E)$, $\Phi(x_i, x_j) \triangleq \exp(\theta_{ij} x_i x_j)$ denote the potential on edge $\{i, j\}$ when $X_i = x_i$ and $X_j = x_j$ and $\Phi(x_A)$ denote the potential due to all edges with both vertices in $A$ when $X_A = x_A$, $\forall A \subseteq V$. In the following, we assume that the girth of the graph is $g > 4$. Consider a node $i$ and a subset of its neighbors $j_1, \cdots, j_k, z$ and a node $w$ which is a neighbor of $z$. We know that the pairwise potentials satisfy $\exp(-\gamma) < \Phi(x_i, x_j) < \exp(\gamma)$. Let $\breve{E} \triangleq E \setminus \{\{i, j_1\}, \cdots, \{i, j_k\}, \{i, z\}, \{z, w\}\}$ and consider the graph $\breve{G}(V, \breve{E})$ with the same potentials on all edges as in $G$. Let $A \triangleq \{i, j_1, \cdots, j_k, z, w\}$ and choose any other set $B \subset V$. Let $P$ and $\breve{P}$ be the probability mass functions corresponding to $G$ and $\breve{G}$ respectively. Similarly let $d(i, j)$ and $\breve{d}(i, j)$

be the distance between $i$ and $j$ in $G$ and $\breve{G}$ respectively. Suppose further that $d(i,B)=d$. Then, $\breve{d}(i,B) > d(A,B) = d$. Note that,

$$\breve{P}(x_A, x_B) = \frac{1}{\breve{Z}} \frac{P(x_A, x_B)}{\Phi(x_A)} \tag{11}$$

where $\breve{Z}$ is an appropriate normalizing constant. Note that $\frac{1}{\breve{Z}} \sum_{x_A} \frac{P(x_A)}{\Phi(x_A)} = \sum_{x_A} \breve{P}(x_A) = 1$. It follows from this that $\exp(-\gamma) < \frac{1}{\breve{Z}} < \exp(\gamma)$. Using (11), the hypothesis that an Ising model has almost exponential correlation decay, we obtain the following inequalities after some algebraic manipulations,

$$|\breve{P}(x_A, x_B) - \breve{P}(x_A)\breve{P}(x_B)| < c2^{D+3} \exp(4\gamma) \exp(-\alpha \min(d, \frac{g-2}{2}))P(x_B) \tag{12}$$

$$\breve{P}(x_B) \geq \exp(-2\gamma) \left(1 - 2^{D+2} c \exp(-\alpha \min(d, \frac{g-2}{2}))\right) P(x_B) \tag{13}$$

$\forall x_A, x_B$.Combining (12) and (13), we obtain

$$|\breve{P}(x_A, x_B) - \breve{P}(x_A)\breve{P}(x_B)| < c2^{D+3} \exp(6\gamma) \frac{\exp(-\alpha \min(d, \frac{g-2}{2}))}{1 - 2^{D+2} c \exp(-\alpha \min(d, \frac{g-2}{2}))} \breve{P}(x_B)$$

and subsequently by marginalizing, we obtain

$$|\breve{P}(x_i, x_B) - \breve{P}(x_i)\breve{P}(x_B)| < c2^{2D+4} \exp(6\gamma) \frac{\exp(-\alpha \min(d, \frac{g-2}{2}))}{1 - 2^{D+2} c \exp(-\alpha \min(d, \frac{g-2}{2}))} \breve{P}(x_B)$$

Let $A' \triangleq A \setminus \{i\}$. Since $d(i, A') = 2$, we have that $\breve{d}(i, A') \geq g - 2$. So, $\exists\ B \subseteq V$ separating $i$ and $A'$ in $\breve{G}$ such that $d(i, B) \geq \frac{g-2}{2}$. Then, $\forall\ x_i, x_{A'}$

$$\begin{aligned} |\breve{P}(x_i \mid x_{A'}) - \breve{P}(x_i)| &= \left| \sum_{x_B} \left( \breve{P}(x_i \mid x_B) - \breve{P}(x_i) \right) \breve{P}(x_B \mid x_{A'}) \right| \\ &< c2^{2D+4} \exp(6\gamma) \tfrac{\exp(-\alpha \frac{g-2}{2})}{1 - 2^{D+2} c \exp(-\alpha \frac{g-2}{2})} \\ &< 2^{-(D+6)} \exp(-2\gamma(D+1)) \left|\sinh(2\beta)\right| \triangleq \breve{\epsilon} \end{aligned} \tag{14}$$

where the last inequality follows from the lower bound on girth $g$ in the hypothesis.

Now consider the graph $\widetilde{G}(V, \widetilde{E})$ where $\widetilde{E} \triangleq \{\{i, j_1\}, \cdots, \{i, j_k\}, \{i, z\}, \{z, w\}\}$. Let the potentials on the edges in $\widetilde{G}$ be the same as those in $G$ and denote the corresponding probability mass function by $\widetilde{P}$. Clearly, we have the following relation between $P, \breve{P}$ and $\widetilde{P}$.

$$P(x_A) = \frac{1}{Z} \breve{P}(x_A) \widetilde{P}(x_A)\ \forall\ x_A$$

where $Z$ is an appropriate normalizing constant. Using (14) and the symmetry of the Ising model (i.e., $\breve{P}(x_i) = \frac{1}{2}$ for $x_i = \pm 1$), we obtain

$$\begin{aligned} P(x_i \mid x_{A'}) &= \tfrac{P(x_i, x_{A'})}{P(x_{A'})} \\ &= \frac{\frac{1}{Z}\breve{P}(x_i, x_{A'})\widetilde{P}(x_i, x_{A'})}{\frac{1}{Z}\sum_{x_i} \breve{P}(x_i, x_{A'})\widetilde{P}(x_i, x_{A'})} \\ &= \frac{\breve{P}(x_i, x_{A'})\widetilde{P}(x_i, x_{A'})}{\sum_{x_i} \breve{P}(x_i \mid x_{A'})\breve{P}(x_{A'})\widetilde{P}(x_i, x_{A'})} \\ &< \frac{\breve{P}(x_i|x_{A'})\widetilde{P}(x_i|x_{A'})}{\left(\frac{1}{2}-\breve{\epsilon}\right)} \\ &< \tfrac{1+2\breve{\epsilon}}{1-2\breve{\epsilon}}\widetilde{P}(x_i \mid x_{A'}) \end{aligned}$$

after some algebraic manipulations. Similarly, we also have

$$P(x_i \mid x_{A'}) > \frac{1-2\breve{\epsilon}}{1+2\breve{\epsilon}}\, \widetilde{P}(x_i \mid x_{A'})$$

which implies

$$\left| P(x_i \mid x_{A'}) - \widetilde{P}(x_i \mid x_{A'}) \right| < 8\breve{\epsilon}$$

Finally, letting $A^* \triangleq A' \setminus \{z\}$, we have,

$$\begin{aligned} & H(X_i \mid X_{A^*}) - H(X_i \mid X_{A'}) \\ &= \sum_{x_{A'}} P(x_{A'}) \sum_{x_i} P(x_i \mid x_{A'}) \log\left(\frac{P(x_i \mid x_{A'})}{P(x_i \mid x_{A^*})}\right) \\ &= \sum_{x_{A'}} P(x_{A'}) D\left(P(X_i \mid x_{A'}) || P(X_i \mid x_{A^*})\right) \\ &\geq \tfrac{1}{2\log 2} \sum_{x_{A'}} P(x_{A'}) \sum_{x_i} |P(x_i \mid x_{A'}) - P(x_i \mid x_{A^*})|^2 \\ &= \tfrac{1}{2} \sum_{x_{A^*}} P(x_{A^*}) \sum_{x_z} P(x_z \mid x_{A^*}) \sum_{x_i} |P(x_i \mid x_{A'}) - P(x_i \mid x_{A^*})|^2 \\ &\geq \tfrac{1}{2} \sum_{x_{A^*}, x_i} P(x_{A^*}) \min_{x_z} P(x_z \mid x_{A^*}) \frac{1}{2} |P(x_i \mid x_{A^*}, x_z = -1) - P(x_i \mid x_{A^*}, x_z = 1)|^2 \\ &\geq \tfrac{1}{4} \sum_{x_{A^*}, x_i} P(x_{A^*}) \frac{\exp(-\gamma D)}{\exp(\gamma D) + \exp(-\gamma D)} \\ &\quad \left(\max\left(0, \left|\widetilde{P}(x_i \mid x_{A^*}, x_z = -1) - \widetilde{P}(x_i \mid x_{A^*}, x_z = 1)\right| - 16\breve{\epsilon}\right)\right)^2 \\ &> \tfrac{1}{8} \sum_{x_{A^*}, x_i} P(x_{A^*}) \exp(-2\gamma D) \left(\frac{|\sinh(2\beta)| \exp(-2\gamma D)}{2} - 16\breve{\epsilon}\right)^2 \\ &> \tfrac{1}{128} \exp(-6\gamma D) \sinh^2(2\beta) \end{aligned}$$

So, we have shown that under the given conditions, an Ising model satisfies (3) with $\epsilon = \frac{1}{128}\exp(-6\gamma D)\sinh^2(2\beta)$. It is straightforward to note that the above proof can also be used to show that the Ising model also satisfies (4) with the same $\epsilon$, completing the proof of the lemma.

■

**Proof** [Theorem 7] The theorem follows directly from Theorem 2, Proposition 4 and Lemma 12. ■